**Travail pratique :**

Ce rapport contient le résumé de toute la durée de travail sur les données de réflectance et acidité d’huile végétale mesurée par un titrage classique Acide/Base à but de construction d’un modèle de prédiction de ce grandeur physicochimique à partir de la réflectance mesurée par un appareil PIR-FTIR (proche infrarouge – méthode de transformée de Fourrier).

Le travail est divisé en trois grandes parties et chacune est subdivisée comme suite :

* Sélection de meilleur procédé de mesure
* Visualisation initiale des données : dans lequel on va identifier les effets expérimentaux d’une manière visuelle
* Identification des effets des anomalies des données : ou on va discuter les tests et les preuves d’existence des effets détectés
* Analyse en composantes principales : cette étape est à but de projeter les données les données dans un sous espace pour les mieux visualiser
* Étude de pertinence des variables : Notre intérêt ici est la discussion de problème de la mauvaise dispersion d’information dans l’espace des variables pour chaque procédé de mesure
* Test de régression : Ce test sert à étudier la qualité des données pour chaque procédé de mesure
* Conclusion : Ou on va choisir d’après les résultats obtenus le procédé de mesure le plus convenable
* Choix de meilleur prétraitement des données
* Lissage de bruit : à but de nettoyer l’effet des anomalies de détecteur d’appareil
* Élimination d’effet de dispersion de la lumière : pour assurer des données propres au niveau d’effet des grandeurs physiques externes
* Conclusion : ou on va choisir le protocole de prétraitement le plus convenable qui donne le maximum possible de performances de régression
* Construction de modèle de prédiction

Et on va finir avec une conclusion générale qui résume le travail avec des proposition d’amélioration pour atteindre l’objectif finale.

**Travail pratique :**

* **Sélection de meilleur procédé de mesure :**

Dans cette partie, on va essayer de choisir quel est le procédé de mesure le plus convenable avec notre méthode d’analyse cible. On sait bien que la mesure de réflectance est le meilleur choix pour l’analyse spectrale de la matière liquide, mais au niveau d’équipements, il y a 3 choix possibles :

* L’installation de deux miroirs derrière la cuve
* 12 janvier
* 31 janvier
* **Visualisation initiale des données :**

Les données d’huile contiennent 1024 colonnes de variables d’entrée : les longueurs d’ondes (en nm) et une colonne additive de variable de sortie : l’acidité (en UA).

A partir de ces donné, on a tracé les spectres bruts qui sont représentés ci-dessous :

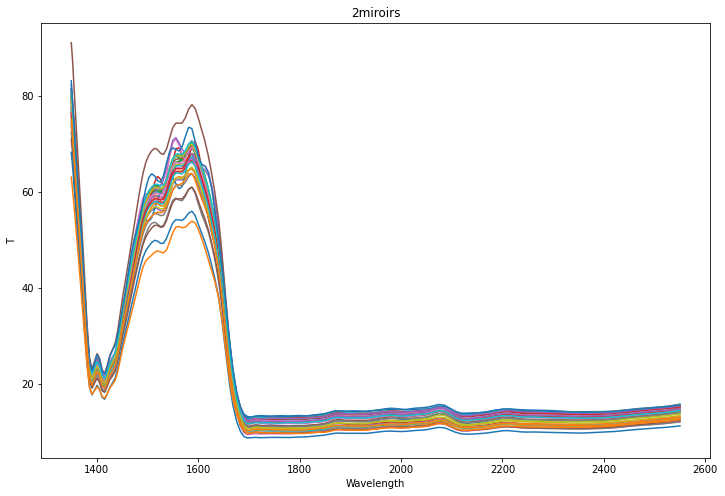


Figure 1 : Spectre brut de procédé de mesure de réflectance avec intégration de deux miroirs

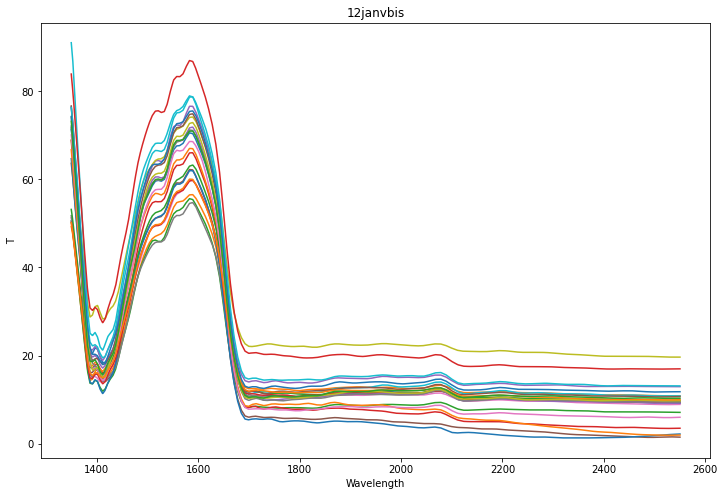


Figure 2 : Spectre brut de procédé de mesure de réflectance 12 janvier

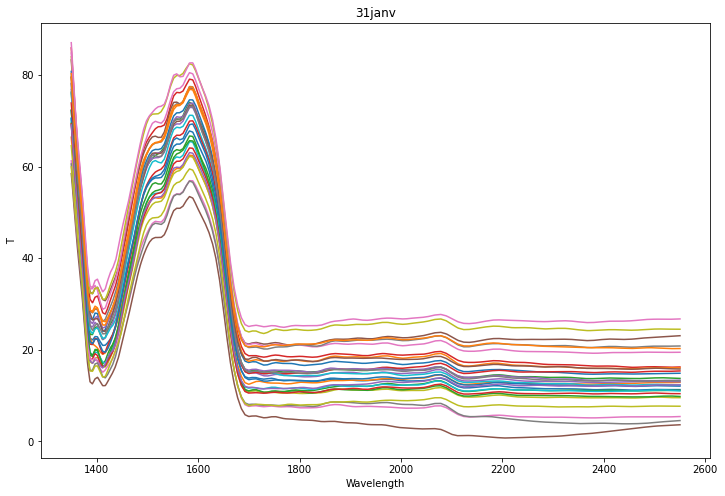
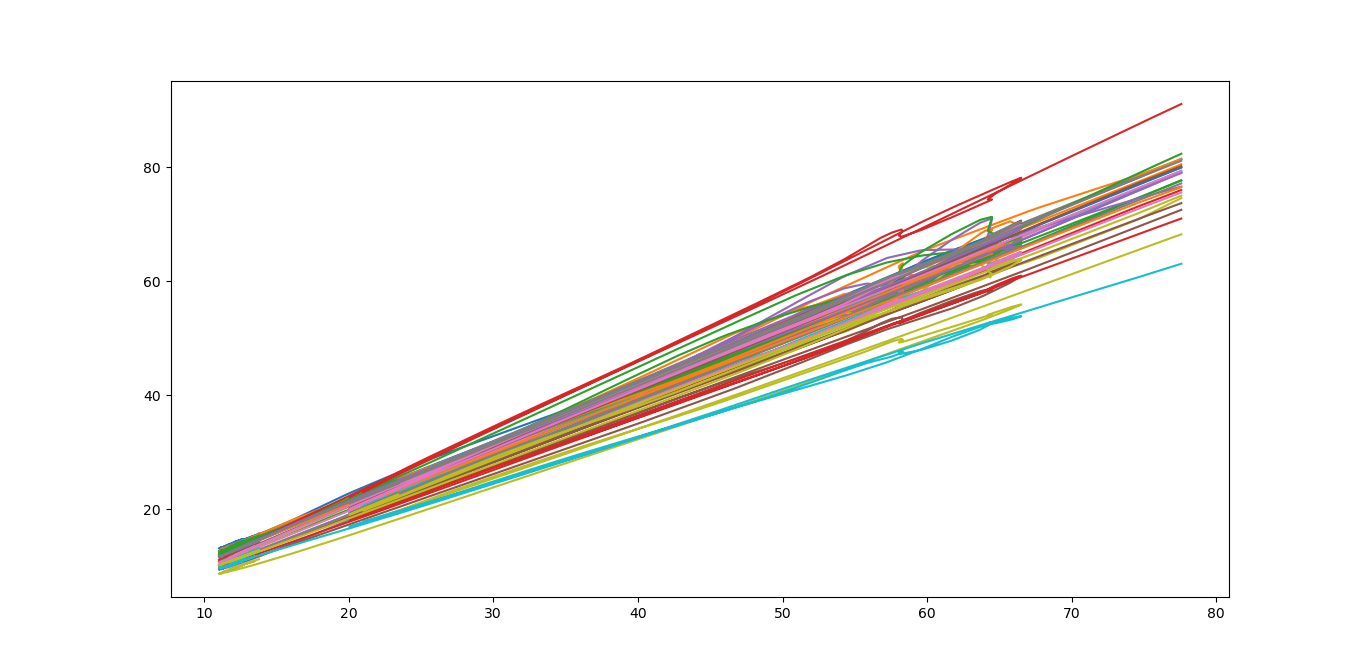


Figure 3 : Spectre brut de procédé de mesure de réflectance 31 janvier

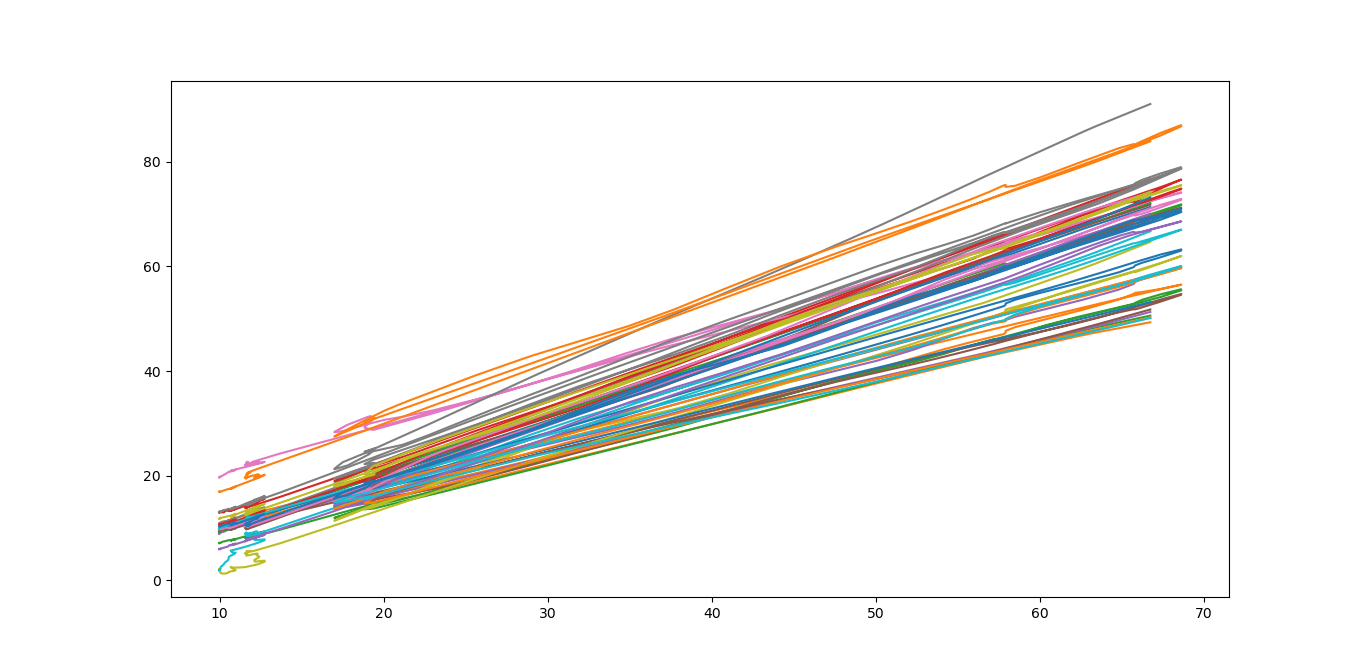
On note l’existence de :

* Effet de bruit de fond pour les 3 procédés de mesure
* Effet de dispersion de lumière faible pour le procédé de mesure avec 2 miroirs et fort pour les autres cas
* **Identification des effets des anomalies des données :**

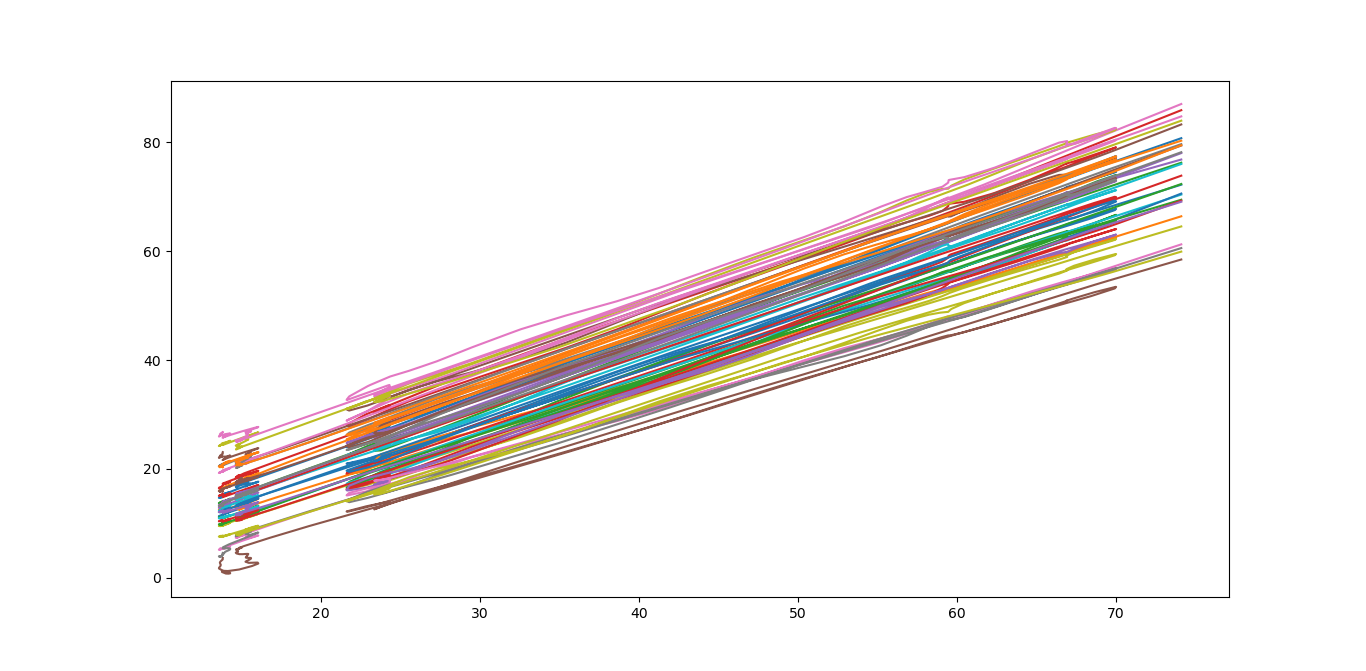
Pour examiner l’effet de ces derniers sur la linéarité de la relation entrée-sortie on va exploiter le plot de spectre moyen contre les spectres des échantillons, et ça donne :



**Figure 4 :** 2 miroirs



**Figure 5 :** 12 janvier



**Figure 6 :** 31 janvier

A partir des 3 plots précédents on note que les anomalies des données extraites à partir des procédés de :

* 12 janvier donnent un effet additif et multiplicatif
* 31 janvier donnent un effet additif
* 2 miroirs donnent un effet multiplicatif
* **Analyse en composantes principales :**

La réduction des dimensions par projection dans un sous-espace donne une meilleure vue pour les données et donne une idée sur les points aberrants, le pourcentage d’information expliquée et le bruit, L’analyse en composantes principales donne :

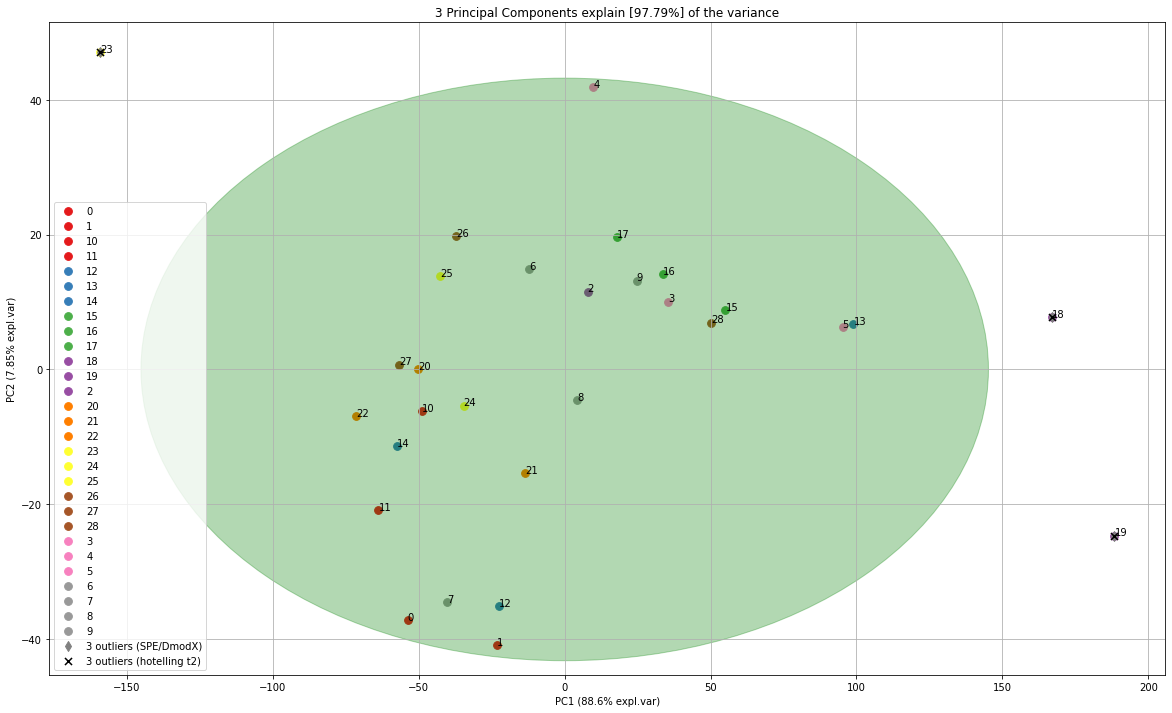


Figure 7 : 2 miroirs 2D

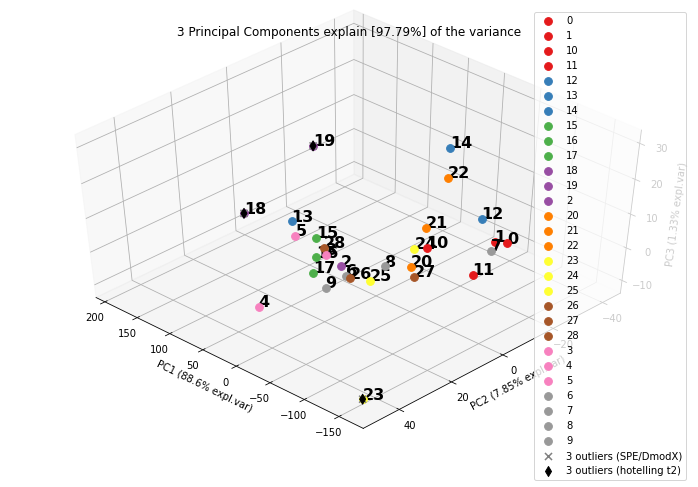


Figure 8 : 2 miroirs 3D

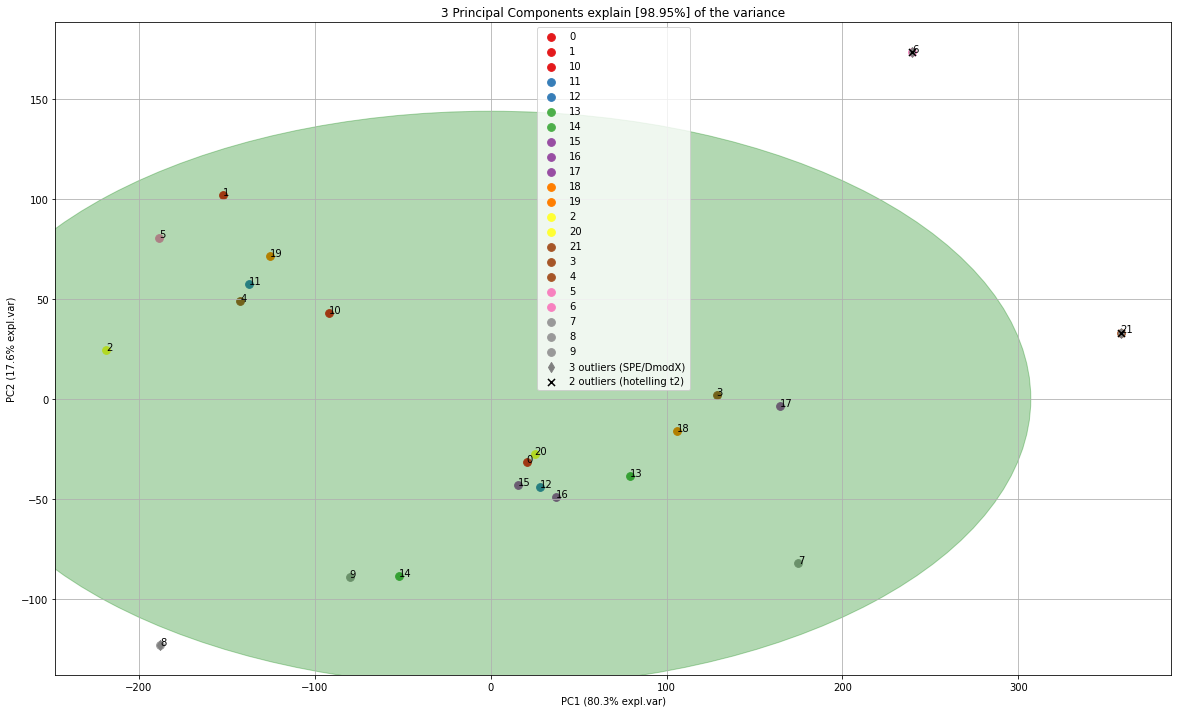


Figure 9 : 12 janvier 2D

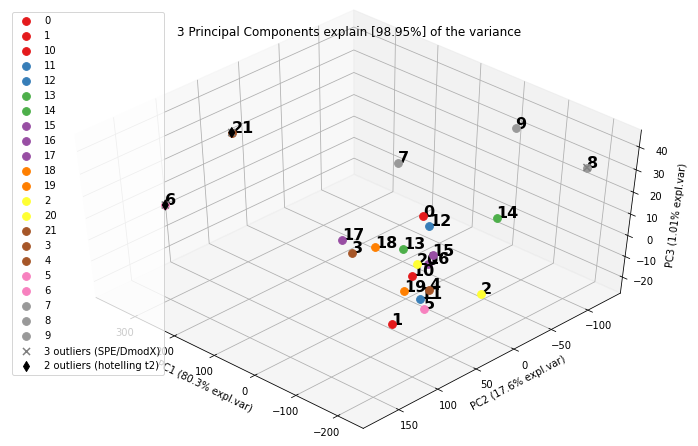


Figure 10 : 12 janvier 3D

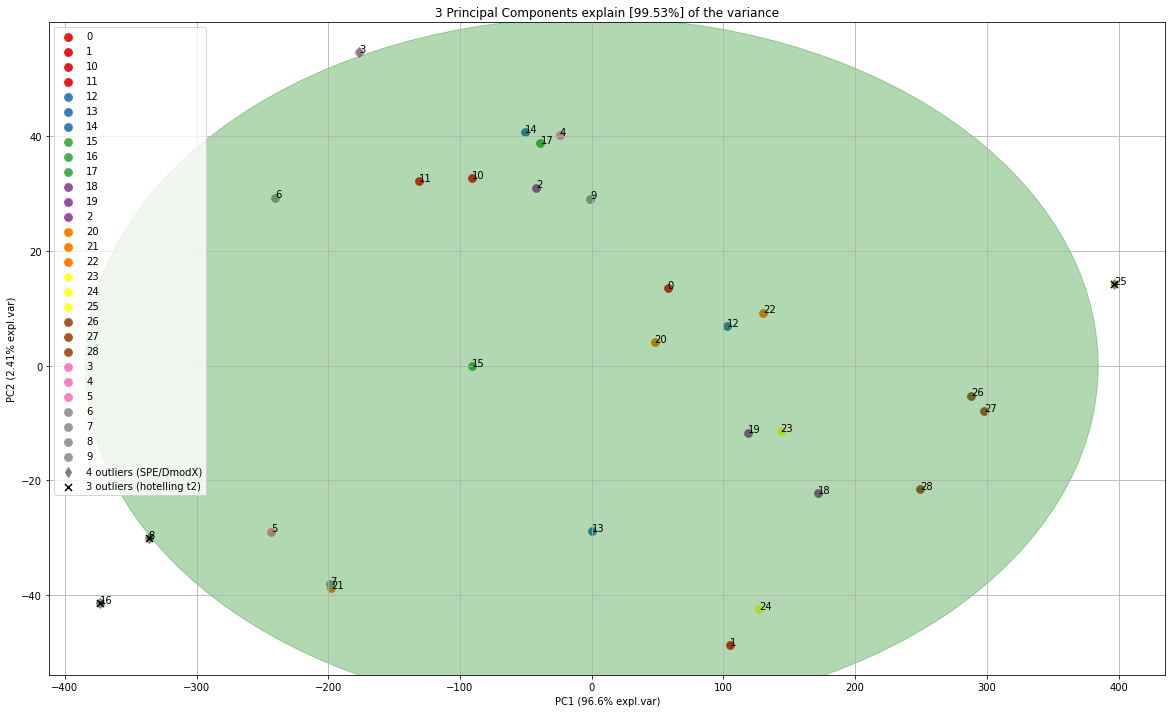


Figure 11 : 31 janvier 2D

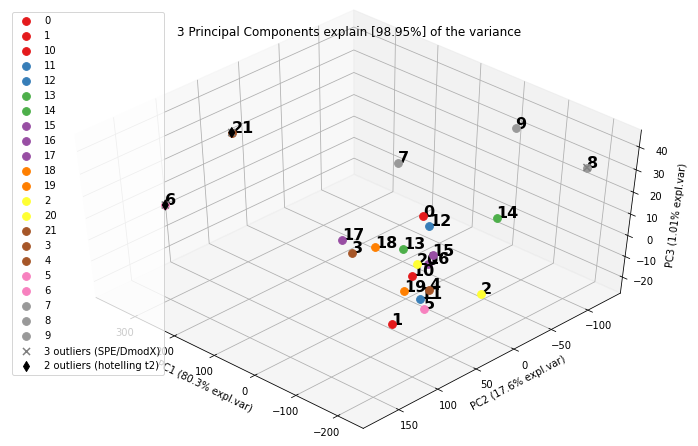


Figure 12 : 31 janvier 3D

On a comme résultats :

|  |  |
| --- | --- |
| Procédé | Nombre les valeurs aberrantes |
| 2 miroirs | 3 |
| 31 janvier | 2 |
| 12 janvier | 2 |

**Tableau 1 :** Nombre de valeurs aberrantes pour chaque projection des données obtenues par les trois procédés de mesure

Mais ce n’est pas suffisant pour décider l’élimination de ces valeurs, il faut d’abord tester la contribution à la création de modèle de prédiction, on va faire cette étape pendant la modélisation des données.

* **Étude de pertinence des variables :**

La forte pertinence entre les variables provoque la modélisation de la même information pour un ensemble des variables, ce qui donne un modèle faible, l’étude de pertinence des variables donne :

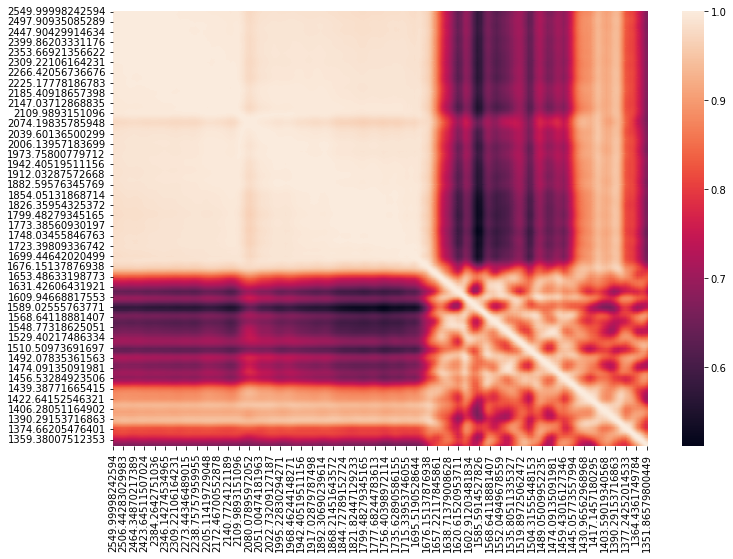


Figure 13 : corrélation des variables de 2 miroirs

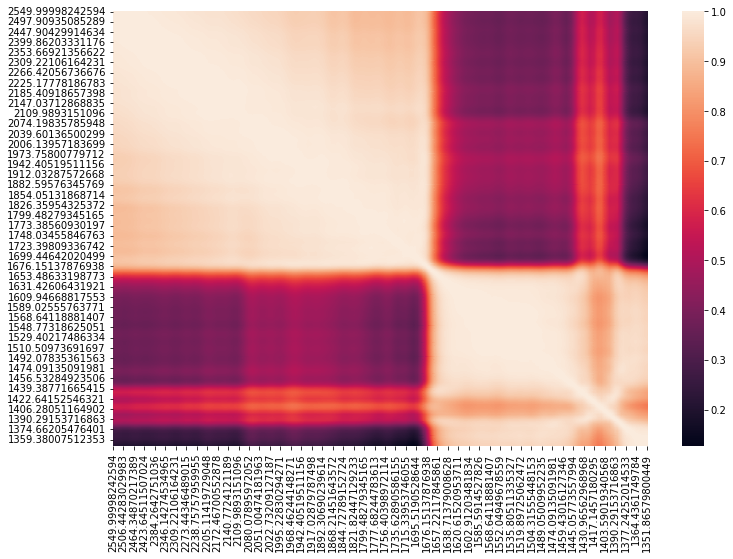


Figure 14 : corrélation des variables de 12 janvier

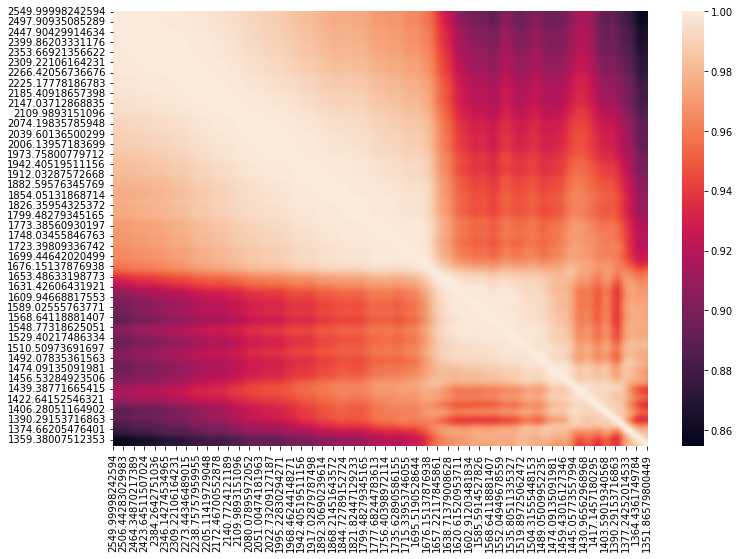


Figure 15 : corrélation des variables de 31 janvier

On peut conclure à partir des trois plots précédents que les données obtenues à partir de procédé de deux miroirs donnent la plus faible pertinence, c’est-à-dire les meilleurs résultats.

* **Test de régression :**

Après plusieurs tests de linéarité et qualité d’apprentissage, prédiction et validation, on a élaboré un modèle de description des données de type PCR (régression sur composantes principales) est ça donne les résultats suivants :

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2 miroirs | 12 Janvier | 31 Janvier |
| **R²** | 0,50983 | 0,123777 | 0,271871 |
| **R²aj** | -0,57944 | -8,63845 | -1,34619 |
| **Fp** | 0,92445 | 1 | 0,999551 |
| **MSEm** | 0,783676 | 0,013924 | 0,040987 |
| **Aic** | 103,3136 | 49,36191 | 47,45213 |
| **Bic** | 130,6595 | 71,18276 | 74,79805 |

**Tableau 2 :** Comparaison des performances pendant la modélisation

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 16 : Diagnostique graphique de 2 miroirs

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 17 : Diagnostique graphique de 12 janvier

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 18 : Diagnostique graphique de 31 janvier

On peut voir clairement à partir des figures précédentes que :

* L’écart à la normalité de résidu est grave et le taux de dépassement des limites de Cook’s dans le cas de procédé de 12 janvier
* Le taux de dépassement des limites de Cook’s est modéré dans le cas de procédé de 31 janvier et 2 miroirs avec une normalité de résidu un peu acceptée mais la distribution est proche de la concavité c’est-à-dire que le modèle est proche d’exiger le terme quadratique pour avoir un meilleur ajustement
* **Conclusion :**

D’après les tests précédents, on peut choisir le procédé de mesure de réflectance basé sur deux miroirs car il a le minimum nombre des points faibles, c’est-à-dire que les données obtenues serrant les plus fortes pendant les prétraitements et la modélisation.

* **Choix de meilleur prétraitement des données :**

Après l’identification des anomalies des données de sortie, on est arrivé à la tache de nettoyage des bruits qui perturbent les données. Il faut respecter les critères statistiques et les protocoles de test pour arriver à la modélisation mathématique la plus exacte comme objectif dans les étapes prochaines.

* **Lissage de bruit :**

Utilisant les algorithmes de lissage, on veut baisser l’effet de bruit d’appareil de mesure de réflectance, on a utilisé deux algorithmes :

* Filtre Savitsky-Golay
* Moving Average

Pour la sélection du meilleur choix, on a utilisé un algorithme de régression linéaire sur composantes principales : PCR, et ça donne :

|  |  |  |
| --- | --- | --- |
|  | **Savitsky-Golay** | **Moving-Average** |
| **R²** | 0.478254 | 0.478363 |
| **R²aj** | -0.681181 | -0.680831 |
| **fp** | 0.952446 | 0.952366 |
| **MSEm** | 0.001009 | 0.001013 |
| **aic** | 103.073455 | 103.067833 |
| **bic** | 130.419372 | 130.413749 |

**Tableau :** Comparaison entre les algorithmes de lissage

On peut conclure que les deux algorithmes sont presque identiques et on peut utiliser un parmi les deux.

* **Élimination d’effet de dispersion de la lumière :**

Utilisant les algorithmes de correction linéaire, on veut baisser l’effet d’augmentation de grandeur optique mesuré par l’appareil (la réflectance), on a utilisé deux algorithmes :

* MSC : Multiplicative Scatter Correction
* SNV : Standard Normal Variation

Les mêmes critères d’élimination seront utilisés ici, et ça donne :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MSC** | **SNV** | **Moving-Average** | **Moving-Average +MSC** | **Moving-Average +SNV** |
| **R²** | 0.498840 | 0.498842 | 0.501743 | 0.499925 | 0.499927 |
| **R²aj** | -0.614849 | -0.614843 | -0.605495 | -0.611353 | -0.611345 |
| **fp** | 0.935336 | 0.935334 | 0.932585 | 0.934318 | 0.934316 |
| **MSEm** | 0.766782 | 0.766785 | 0.771244 | 0.768450 | 0.768454 |
| **aic** | 103.956626 | 103.956522 | 103.788142 | 103.893770 | 103.893623 |
| **bic** | 131.302542 | 131.302438 | 131.134059 | 131.239686 | 131.239540 |

**Tableau :** Comparaison entre les algorithmes de correction linéaire

D’après les chiffres de tableau, on peut conclure qu’il n’y a pas d’intérêt de mettre les deux types de correction à la fois car ils donnent les mêmes résultats, c’est-à-dire que l’appareil n’est pas sensible à la dispersion de la lumière.

* **Conclusion :**

D’après les analyses précédentes, en peut se contenter avec l’élimination de bruits d’appareil pour assurer la prise de maximum possible des composantes principales pendant l’étape de la modélisation, ce qui implique la conservation de maximum possible des informations utiles.

* **Construction de modèle de prédiction :**

Après plusieurs algorithmes de modélisation, on a choisi les deux les plus convenables pour atteindre l’objectif :

* PCR : Régression sur composantes principales
* PLS : Régression des moindres carrés partiels

La base de données sera devisée en deux parties d’une façon aléatoire :

* Une partie de 80% de la base sera utilisée pour la calibration
* La 20% restante sera utilisée pour le test de modèle

Cette division aléatoire va provoquer une erreur d’échantillonnage, pour cette raison on va faire une répétition d’échantillonnage plusieurs fois jusqu’à l’obtention des résultats favorables.

Les critères de performances de modèle utilisés dans cette opération sont :

* Le coefficient de détermination de calibration
* Le coefficient de détermination de cross validation
* Le coefficient de détermination de test
* La racine d’erreur moyenne carrée de calibration
* La racine d’erreur moyenne carrée de cross validation
* La racine d’erreur moyenne carrée de test

Ce protocole donne comme résultats :

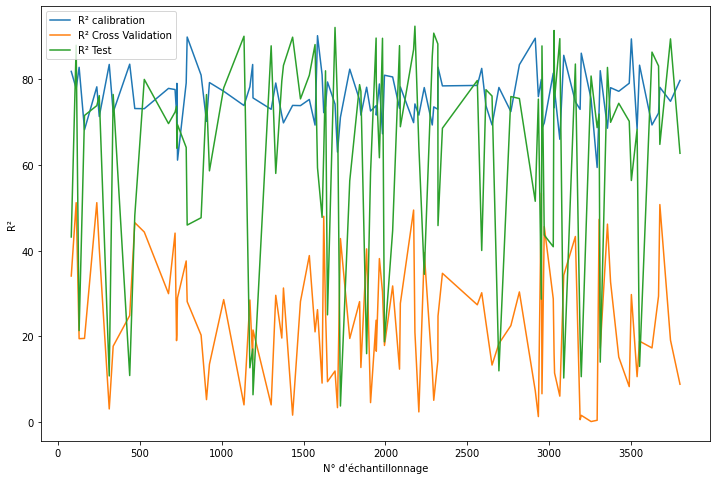


Figure : L’évolution de coefficient de détermination R² en fonction d’échantillonnage

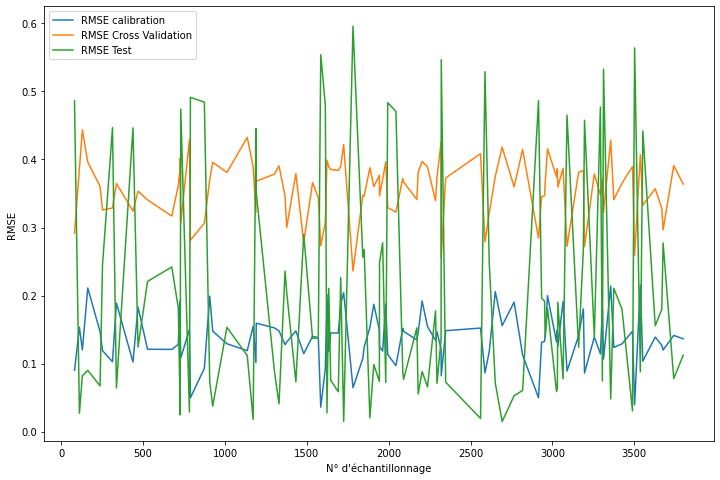


Figure : L’évolution de La racine d’erreur moyenne carrée RMSE en fonction d’échantillonnage